AD-A051 557 TRACOR INC AUSTIN TEX F/G 17/1

CONFIDENCE INTERVALS ASSOCIATED WITH RANDOM SIGNAL PROCESSING.(U)

DEC 68 H D RECORD, H A REEDER

UNCLASSIFIED NL

| OF |

AD
A051557

END
DATE
FILMED
4-78
DDC

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

TRACOR Project 072 951

11 31 December 1968

9 MEMORANDUM rept.

12 39P.

1

B.S.

TO: J. J. Dow

~~FROM~~: 10 H. D. Record [redacted] H. A. Reeder

~~SUBJ~~: 6 CONFIDENCE INTERVALS ASSOCIATED WITH RANDOM SIGNAL PROCESSING.

The enclosed note contains the necessary procedures for determining confidence intervals for several types of estimates that occur frequently in sonar signal analysis. A table of contents is given below.

H. D. Record

H. D. Record

H. A. Reeder

H. A. Reeder

RR/ca

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (continued)

LIST OF TABLES

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

1. INTRODUCTION

When a random signal is sampled and digitally analyzed, the results of the analysis are estimates of the true parameters or statistics describing the random signal. In many cases there exists difficulty in assigning a level of confidence to such an estimate. That is, the estimate can never be said to equal the true parameter value, but can be said to lie within specified confidence limits, or error bars, with a given probability. The confidence limits are highly dependent upon the characteristics of the particular random signal being analyzed and the parameter being estimated.

The most important single characteristic is the number of independent samples available from the random signal. For a random signal which contains both amplitude and phase information, the number of independent samples available is

$$N_I = 2\,BT \quad \text{(amplitude } \underline{\text{and}} \text{ phase samples)}$$

where

N_I - is the number of independent samples;

B - is the bandwidth of the random signal; and

T - is the signal duration, (the length of the sampled function).

In a random process which contains only amplitude information, such as at the output of a detector-averager, half the independent samples have been destroyed,

$$N_I = BT \quad \text{(amplitude } \underline{\text{or}} \text{ phase samples).}$$

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

This memorandum contains the necessary procedures for determining confidence intervals for several types of estimates that occur frequently in sonar signal analysis. Covered in Section 2 are confidence intervals for the sample mean and sample variance from a normal population. Section 3 contains a description of the chi-square goodness-of-fit test, used to test the equivalence of a measured probability density function for sampled data to some hypothesized density function. Section 4 describes the Kolmogorov (K) statistic, used to set a confidence band about an entire probability distribution function; and a binomial statistic, used to set confidence limits about each point on an empirical distribution function. References are listed in Section 5.

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

2. PARAMETER ESTIMATION

2.1 SAMPLE MEAN

The sample mean of a random signal is defined as

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i,$$

where

$\bar{x}$ - is the sample mean;

N - is the number of samples analyzed; and

x_i - is the value of the i^{th} sample.

Note that the x_i are random variables described by some type of statistics. The sample mean is formed by a linear combination of the x_i, thus $\bar{x}$ is also a random variable and may only be considered as an estimate of the true mean, μ_x; that is,

$$\bar{x} = \hat{\mu}_x,$$

where the "$\wedge$" denotes "estimate."

A confidence interval for the mean value estimate is then some interval about the sample mean within which the true mean value lies with a given probability, on the average. For the purposes of Section 2 of this memorandum, the x_i will be assumed to be samples of a normal random variable. A confidence interval for the sample mean of a normal population is then [1, page 140]

$$(1-\alpha) = \text{Prob}\left\{\left(\bar{x} - \frac{st_{n\,;\,\alpha/2}}{\sqrt{N_I}}\right) \le \mu_x \le \left(\bar{x} + \frac{st_{n\,;\,\alpha/2}}{\sqrt{N_I}}\right)\right\},$$

where

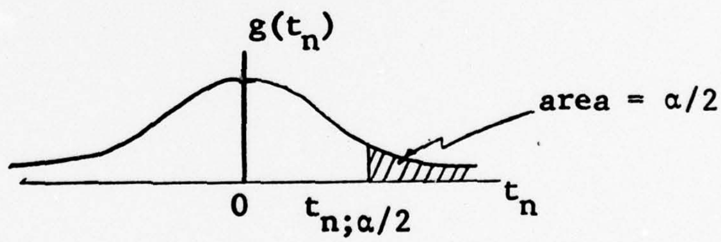
$(1-\alpha)$ - is the probability that μ_x will lie within the confidence interval;

s - is the unbiased sample standard deviation defined in Section 2.1;

N_I - is the number of independent samples;

n - is the number of degrees of freedom, ($n=N_I-1$ for the sample mean); and

$t_{n;\,\alpha/2}$ - is the value of the variable, t_n, in Student's t distribution with n degrees of freedom, such that $\text{Prob}(t_n > t_{n;\alpha/2}) = \alpha/2$. (The Student t density is symmetric about $t_n = 0$.)



Tables of the Student t distribution may be used to calculate this interval, or the curves in Fig. 1 may be used. These curves are plots of $\frac{t_{n;\,\alpha/2}}{\sqrt{N_I}}$ versus n for different values of α. It should be emphasized at this point that the number of <u>independent</u> samples N_I, is necessary for this calculation no matter

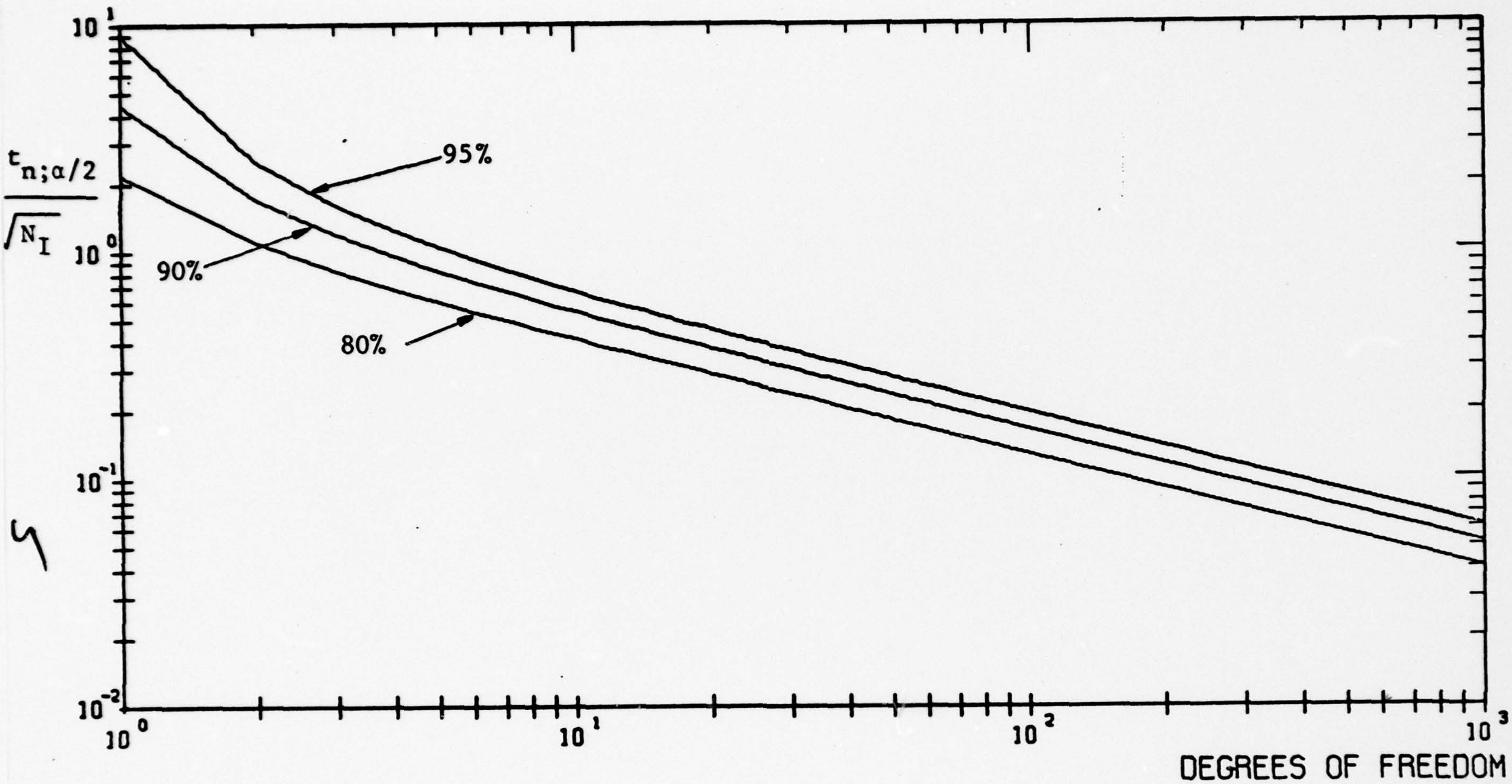


Fig. 1 NORMALIZED CONFIDENCE INTERVAL FOR GAUSSIAN SAMPLE MEAN

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

how great the sampling rate. The curves in Fig. 1 arise from the assumption that the samples x_i are from a normal population, however, for $N_I > 10$ the sampling distribution for $\overline{x}$ approaches a normal distribution [1, page 136].

2.2 SAMPLE VARIANCE

The sample variance of a random signal is defined as

$$s^2 = \left(\frac{N_I}{N_I - 1}\right) \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

$$s^2 = \hat{\sigma}_x^2,$$

where

s^2 - is the sample variance (unbiased estimate);

N_I - is the number of independent samples;

N - is the number of samples analyzed;

x_i - is the value of the i^{th} sample; and

σ_x^2 - is the true variance.

This definition of the sample variance arises from the unbiased estimate in which only independent samples are taken [1pages 125,126].

Again, assuming that the x_i are samples from a normal population, a confidence interval for the sample variance may be defined as [1 page 140],

$$(1-\alpha) = \text{Prob}\left\{\frac{ns^2}{\chi^2_{n;\alpha/2}} \le \sigma_x^2 \le \frac{ns^2}{\chi^2_{n;1-\alpha/2}}\right\},$$

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

where

$\chi^2_{n;\beta}$ is the value of the chi-square variable with n degrees of freedom, χ^2_n, such that $\text{Prob}\left\{\chi^2_n > \chi^2_{n;\beta}\right\} = \beta$; and

$n = (N_I - 1)$ for the sample variance.

Tables of the chi-square distribution function may be used to calculate the above interval, or the curves in Figs. 2 and 3 may be used. These curves are plots of $\frac{n}{\chi^2_{n;\alpha/2}}$ and $\frac{n}{\chi^2_{n;1-\alpha/2}}$ versus n for several values of α. Table I includes values of $\frac{n}{\chi^2_{n;1-\alpha/2}}$ from Fig. 3 for $1 \le n \le 5$, since some values were too large to be conveniently plotted.

Once again it is emphasized that the above confidence interval depends upon the number of degrees of freedom, n, no matter how great the sampling rate. A high sampling rate allows easier and more detailed practical reconstruction of the original analog waveform but does not effect theoretically the confidence limits.

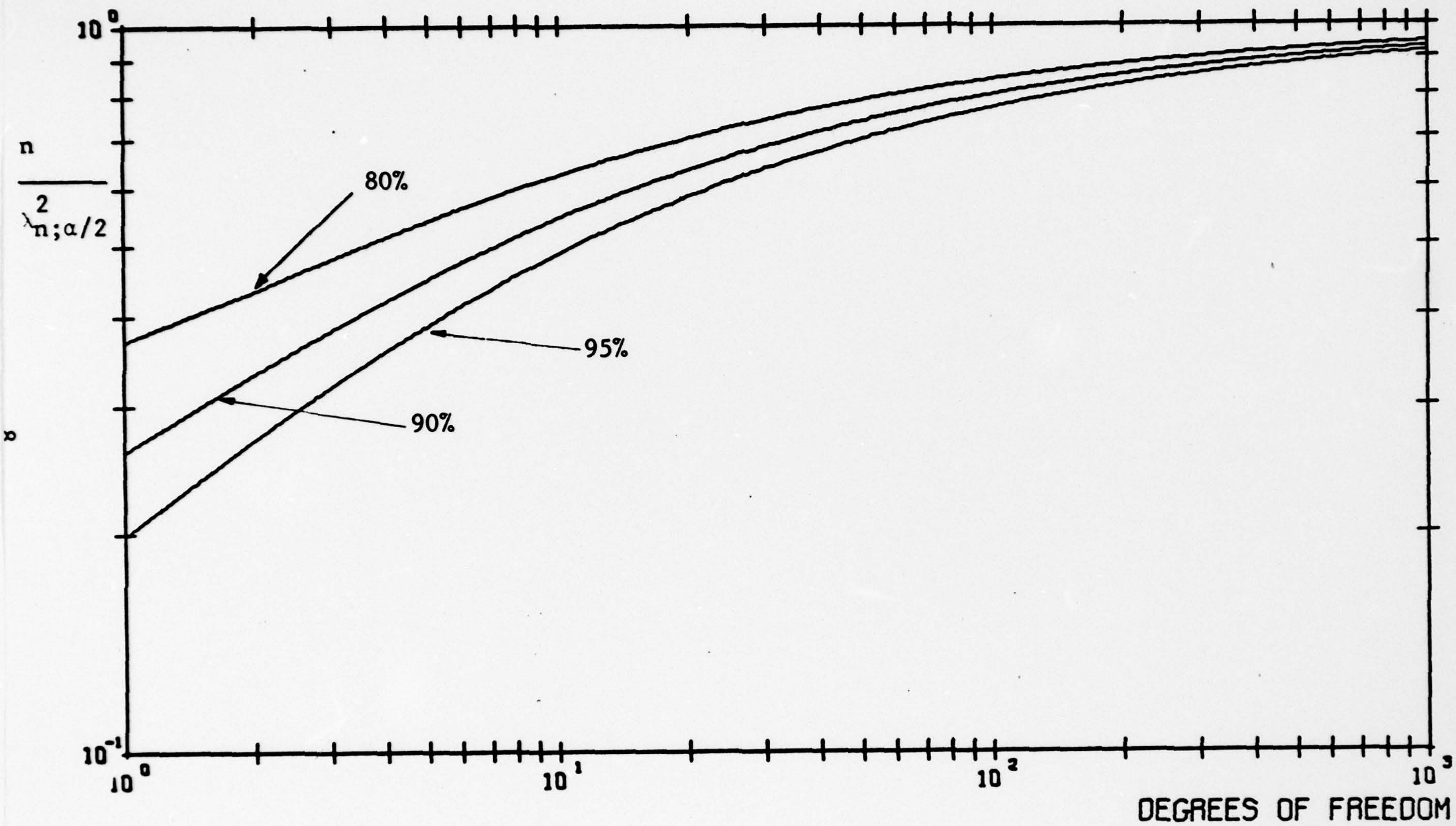


Fig. 2 NORMALIZED LOWER CONFIDENCE LIMITS FOR GAUSSIAN SAMPLE VARIANCE

TRACOR

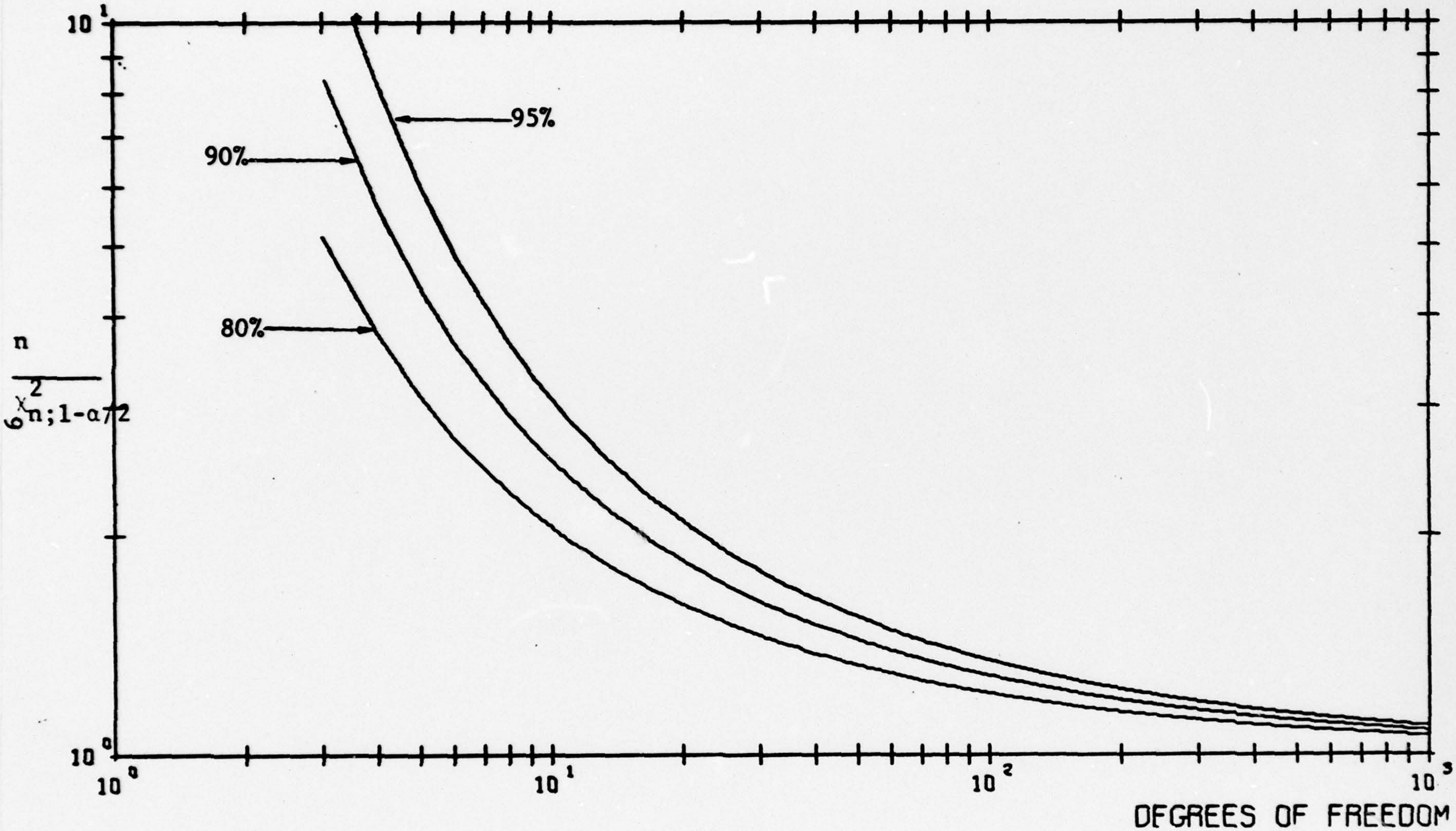


Fig. 3 NORMALIZED UPPER CONFIDENCE LIMITS FOR GAUSSIAN SAMPLE VARIANCE

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

TABLE I NORMALIZED UPPER CONFIDENCE LIMITS FOR GAUSSIAN SAMPLE VARIANCE, $(1 \le n \le 5)$.

Degrees of Freedom, n	$\frac{n}{\chi^2_{n;0.90}}$	$\frac{n}{\chi^2_{n;0.95}}$	$\frac{n}{\chi^2_{n;0.975}}$
1	63.82	257.21	1018.26
2	9.49	19.50	39.50
3	5.13	8.53	13.90
4	3.76	5.63	8.26
5	3.11	4.36	6.02

EXAMPLE

Consider a random time function described by Gaussian statistics which has a duration of one second with a known bandwidth of 100 Hz. The sample mean is found to be 1.0, and the sample variance is found to be 4.0, that is

$$\bar{x} = 1.0 \text{ and}$$

$$s^2 = 4.0.$$

The number of independent samples, N_I is

$$N_I = 2\ BT$$

$$= 2(100)(1)$$

$$N_I = 200 \text{ independent samples.}$$

Determine a 90% confidence interval for the true mean value μ_x and the true variance σ_x^2.

For the mean value confidence interval, we have the relationship on page 4, which gives

$$0.9 = \text{Prob}\left\{1.0-2.0\left(\frac{t_{199;\ 0.45}}{\sqrt{200}}\right) \le \mu_x \le 1.0-2.0\left(\frac{t_{199;\ 0.45}}{\sqrt{200}}\right)\right\} .$$

Thus, given $n = N_I - 1 = 199$ degrees of freedom, the appropriate $\frac{t_{n;\ \alpha/2}}{\sqrt{N_I}}$ value comes from the ordinate of the 90% curve at abscissa n = 199. This value is approximately 0.117. Thus a 90% confidence statement about the true mean value is

$$0.9 = \text{Prob}\left\{1.0-(2.0)(0.117) \le \mu_x \le 1.0 + (2.0)(0.117)\right\}$$

$$0.9 = \text{Prob}\left\{0.766 \le \mu_x \le 1.234\ .\right\}$$

That is, there is 90% confidence that the true mean μ_x will lie between 0.766 and 1.234 when $\overline{X} = 1.0$, $s^2 = 4.0$, and $N_I = 200$.

For the variance confidence interval, we have the relationship on page 6, which gives

$$0.9 = \text{Prob}\left\{\left(\frac{199}{\chi^2_{199;0.05}}\right)(4.0) \le \sigma_x^2 \le \left(\frac{199}{\chi^2_{199;\ 0.95}}\right)(4.0)\right\}.$$

Thus for n = 199 degrees of freedom and a 90% confidence interval, Figure 2 gives $\frac{199}{\chi^2_{199;\ 0.05}}$ as approximately 0.86, and Figure 3 gives $\frac{199}{\chi^2_{199;\ 0.95}}$ as approximately 1.2. Thus a 90% confidence

statement for the true variance is

$$0.9 = \text{Prob}\left\{(0.86)(4.0) \le \sigma_x^2 \le (1.2)(4.0)\right\}$$

$$0.9 = \text{Prob}\left\{3.44 \le \sigma_x^2 \le 4.8\right\} .$$

That is, there is 90% confidence that the true variance will lie between 3.44 and 4.8 when $s^2 = 4.0$ and $N_I = 200$.

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

3. SAMPLED DATA PROBABILITY DENSITY FUNCTION

This section discusses a method by which one may test the equivalence of a measured probability density function for sampled data to some hypothesized probability density function. The method is known as the chi-square goodness-of-fit test [1 pages 146, 147]. Once a hypothetical density function is chosen, this test can be used to determine whether the overall measured sampled-data density function approaches the hypothetical density.

The general procedure of the chi-square goodness-of-fit test involves the use of a statistic with an approximate chi-square distribution as a measure of discrepency between the observed and hypothetical densities. Consider a series of samples $\{x_i\}$, i = 1, N taken from a random function, in which there are N_I independent samples. Group the samples into K bins to form a frequency histogram or discrete observed density function.

Denote the number of independent samples falling within the i^{th} bin as f_i. Calculate the number of independent samples expected to fall within the i^{th} bin if indeed the hypothetical density were correct, and denote this number by F_i, given by

$$F_i = N_I \int_{\lambda_i}^{\lambda_{i+1}} p_h(\lambda)\, d\lambda,$$

where

N_I	- is the number of independent samples from the random function;
$p_h(\lambda)$	- is the hypothetical probability density function; and
$(\lambda_i, \lambda_{i+1})$	- is the interval describing the i^{th} bin.

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

The discrepancy between observed frequency and expected frequency for the i^{th} bin is then $(f_i - F_i)$, which may be used to find a sample statistic X^2, given by

$$X^2 = \sum_{i=1}^{K} \frac{(f_i - F_i)^2}{F_i} .$$

The distribution for X^2 is approximately the same as a chi-square (χ^2_n) distribution, where n, the number of degrees of freedom is equal to K minus the number of independent restrictions on the observations. In general, [2, page 177],

$$n = K-1-b ,$$

where b is the number of parameters in the population description determined from the random sample. The normal distribution is completely characterized by the mean and standard deviation, so

$$n = K-3 \quad \text{(Normal Distribution)}.$$

For Rayleigh statistics only one parameter is necessary, so

$$n = K-2 \text{ (Rayleigh Distribution)}.$$

Once the quantity X^2 has been computed it may be tested for goodness-of-fit. The region of acceptance is such that

$$X^2 \leq \chi^2_{n;\alpha} .$$

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

That is, if the value of X^2 is less than or equal to the appropriate chi-square variate, then the probability is at least $(1-\alpha)$ that the samples under consideration are described by the assumed theoretical density. Or, conversely the probability of a sample deviating from the assumed density is α.

In order to apply the chi-square goodness-of-fit test, the number of bins, K must be chosen with care [1 page 147]. Table II describes the minimum number of bins for N independent samples and $\alpha = 0.05$.

TABLE II MINIMUM OPTIMUM NUMBER OF BINS (K) FOR N_I INDEPENDENT SAMPLES AND $\alpha = 0.05$.

N_I	200	400	600	800	1000	1500	2000
K	16	20	24	27	30	35	39

In addition to the stipulation in Table II, the chi-square goodness-of-fit test works best when N_I is large enough to assure that at least 10 samples fall into each bin, especially near the tails of the population density.

The following procedure is a summary of the chi-square goodness-of-fit test for the case when the population distribution is not known and the number of independent samples and bins have been chosen correctly.

1. Hypothesis H: The data $\{x_i\}$ are a sample of a random variable with density $p_h(\lambda)$.

2. Purpose of test: To determine whether the data $\{x_i\}$ may be considered as consistent with hypothesis H.

3. Steps in test:

(a) Form the statistic $X^2 = \sum_{i=1}^{K} \frac{(f_i - F_i)^2}{F_i}$.

(b) Determine the number of degrees of freedom n.

(c) Select a level of significance $\alpha = 0.01, 0.05, \ldots$).

(d) From tables of the χ_n^2 distribution determine $\chi^2_{n;\alpha}$ such that $\text{Prob}(\chi_n^2 \geq \chi^2_{n;\alpha}) \cong \alpha$.

(e) If the test statistic X^2 is greater than $\chi^2_{n;\alpha}$, then the hypothesis H is <u>rejected</u> at the α level of significance.

(f) If the test statistic X^2 is such that $X^2 \leq \chi_{n;\alpha}$, then the sample function may be considered as consistent with hypothesis H.

It should be observed at this point that for n > 30, the χ_n^2 distribution may be approximated by a normal distribution with mean n and standard deviation $\sqrt{2n}$. Also, the variable $\sqrt{2\chi^2_n}$ is approximately normally distribution with mean $\sqrt{2n-1}$ and unit standard deviation.

4. SAMPLED DATA PROBABILITY DISTRIBUTION FUNCTION

4.1 KOLMOGOROV STATISTIC

Often it is necessary or more convenient to obtain a probability distribution function estimate for sampled random data. The method of obtaining a confidence level described in this section allows a confidence band to be placed on such a distribution estimate. The method, known as the Kolmogorov (K) statistic [3 page 452], is an alternative to the chi-square test described in Section 3. The chi-square goodness-of-fit test gives a general idea about the variation of a sampled-data density function varies about a hypothetical density. The K statistic allows one to determine a confidence band about the distribution function estimate so that a confidence statement can be made associated with a given probability.

In order to use the K statistic, the actual distribution function describing the observed random process must be continuous. Actual expressions for the K statistic are quite complicated, however, asymptotic approximations are available and are listed in Table III. The approximations are conservative for all values of N_I and satisfactory for $N_I \geq 80$. The approximations are asymptotic in that they approach asymptotically the true K statistic as N_I increases.

Let $\hat{F}(x)$ be the estimate of the true probability distribution function $F(x)$. Then $F(x)$ may be assigned the confidence interval

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

$$\text{Prob}\left\{\left|\hat{F}(x) - d\beta\right| \leq F(x) \leq \left|\hat{F}(x) + d\beta\right|\right\} = \beta$$

for all x, where β is the desired confidence level and $(\pm d\beta)$ is the interval about $\hat{F}(x)$ within which F(x) will fall 100 β% of the time, on the average. For example using Table III, if $N_I = 100$, then for any x, F(x) will be between $(\hat{F}(x) - 0.13581)$ and $(F(x) + 0.13581)$ 95% of the time, on the average.

TABLE III

KOLMOGOROV CONFIDENCE INTERVAL

Confidence Level, β	Confidence Interval, $d\beta$
0.95	$\frac{1.3581}{\sqrt{N_I}}$
0.99	$\frac{1.6276}{\sqrt{N_I}}$

4.2 BINOMIAL STATISTIC

The K statistic just described gives a uniform bound on the probability distribution function which may make the confidence interval somewhat wide in some cases. An alternative is to make the confidence limits dependent upon the observed distribution. This may be accomplished by setting a threshold for the input random signal and counting the number of samples falling above and below this threshold. It is here that the binomial distribution is introduced. Note that this is essentially the same as the sorting procedure used to find a sampled data probability distribution function.

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

The probability that a sample will fall below a threshold x_o is $F(x_o)$, while the sorting procedure yields $\hat{F}(x_o)$. Using the binomial distribution, it is possible to set confidence intervals about $F(x_o)$. Mathematical details of this process will be the subject of a forthcoming memorandum.

The results of this method are given in Figs. 4 through 17. Each figure includes curves for a given N_I and several confidence levels. For each level, confidence limits are plotted for $\hat{F}(x)$ in terms of the true value $F(x)$. The curves may be utilized in the following manner. For a value of $F(x)$ obtained with N_I independent samples, choose the proper figure and confidence level. Draw a line parallel to the abscissa ($F(x)$), with ordinate ($\hat{F}(x)$). This line will intersect the proper curves for the chosen level of confidence. The values of $F(x)$ at which the straight line intersects the proper curves give the interval about $\hat{F}(x)$ within which $F(x)$ will lie with the given confidence. For $N_I \geq 20$, logarithmic curves are presented to give more detail about the tails of the distributions. Note the symmetry present in the curves.

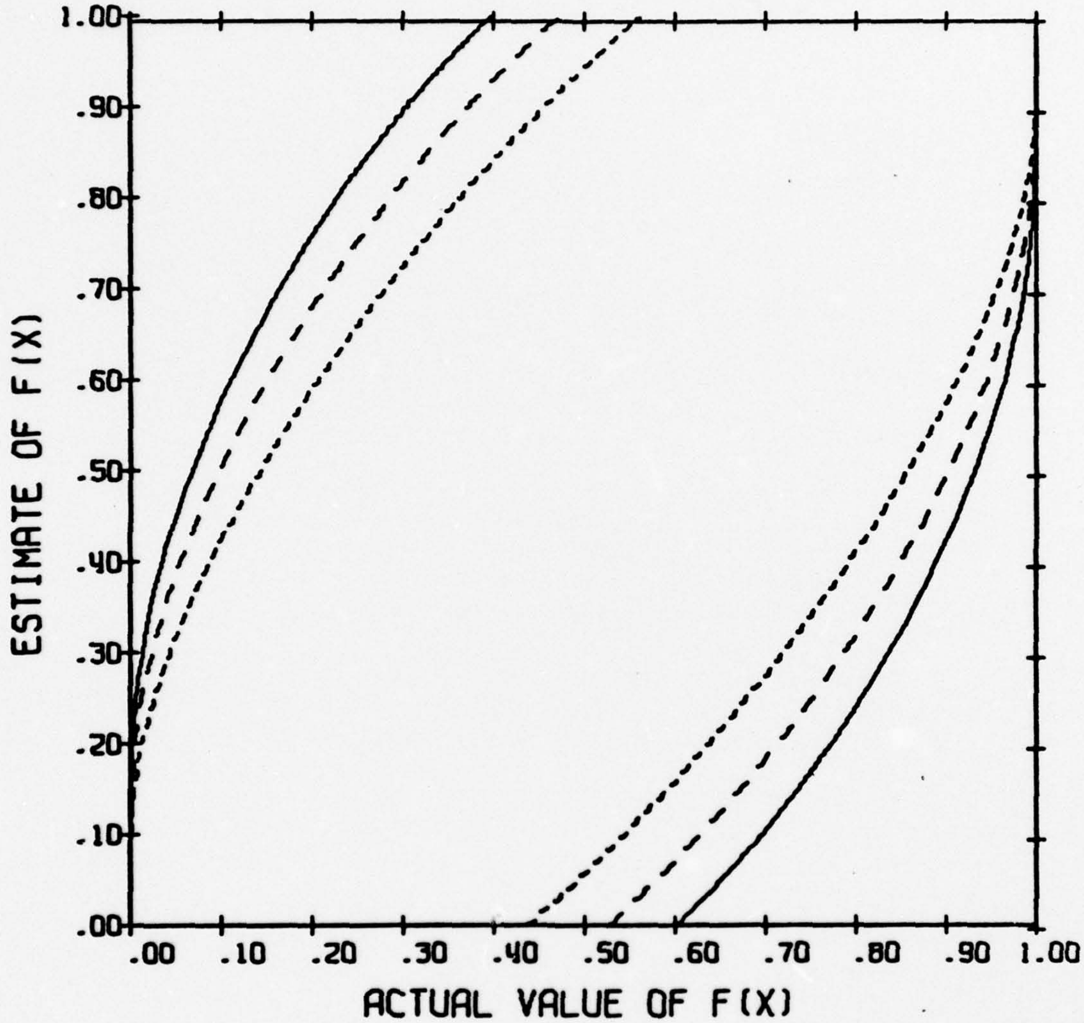


Fig. 4 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 4
——.95 - - -.90 -----.80

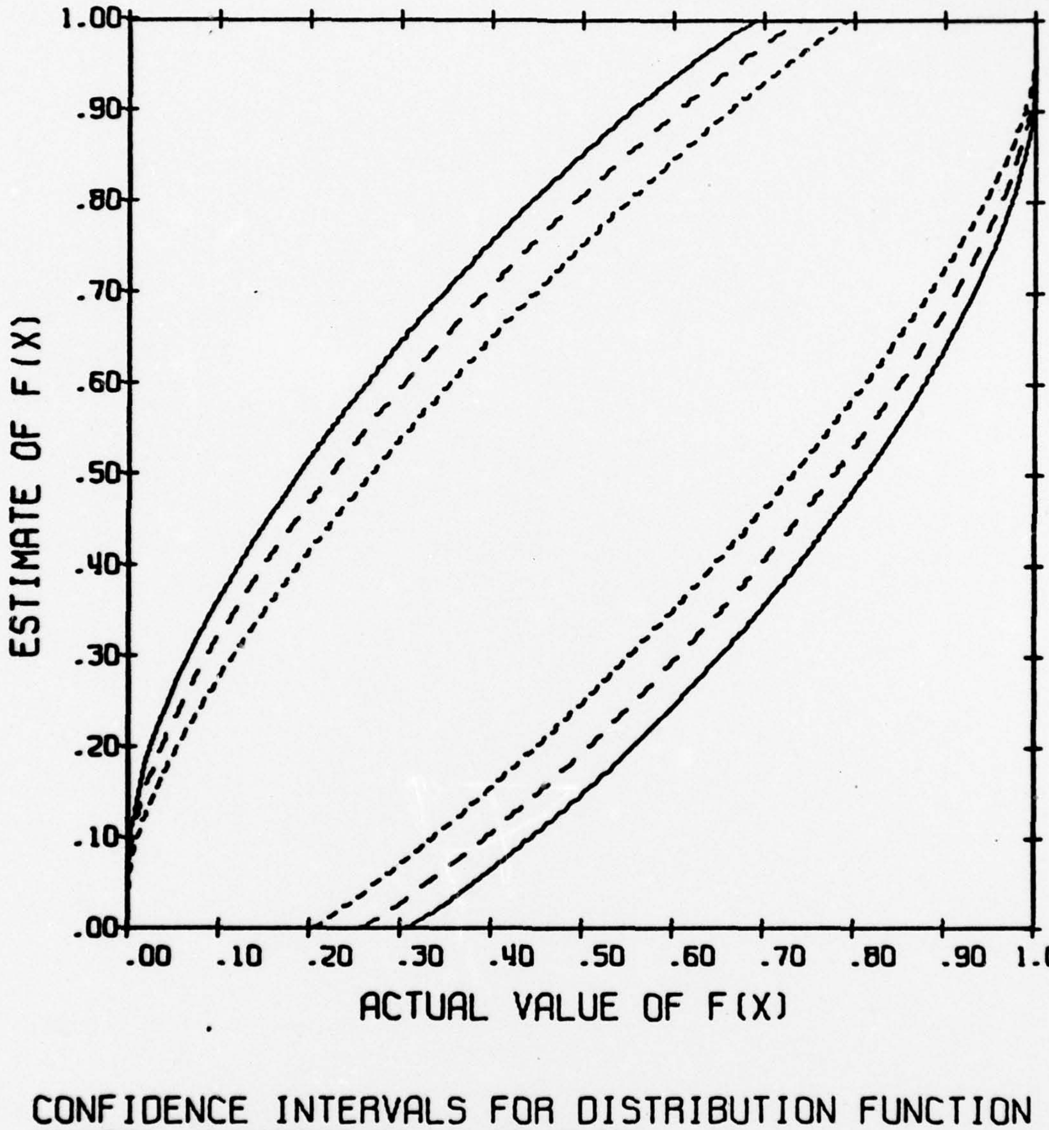


Fig. 5 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 10
——.95 - - -.90 -----.80

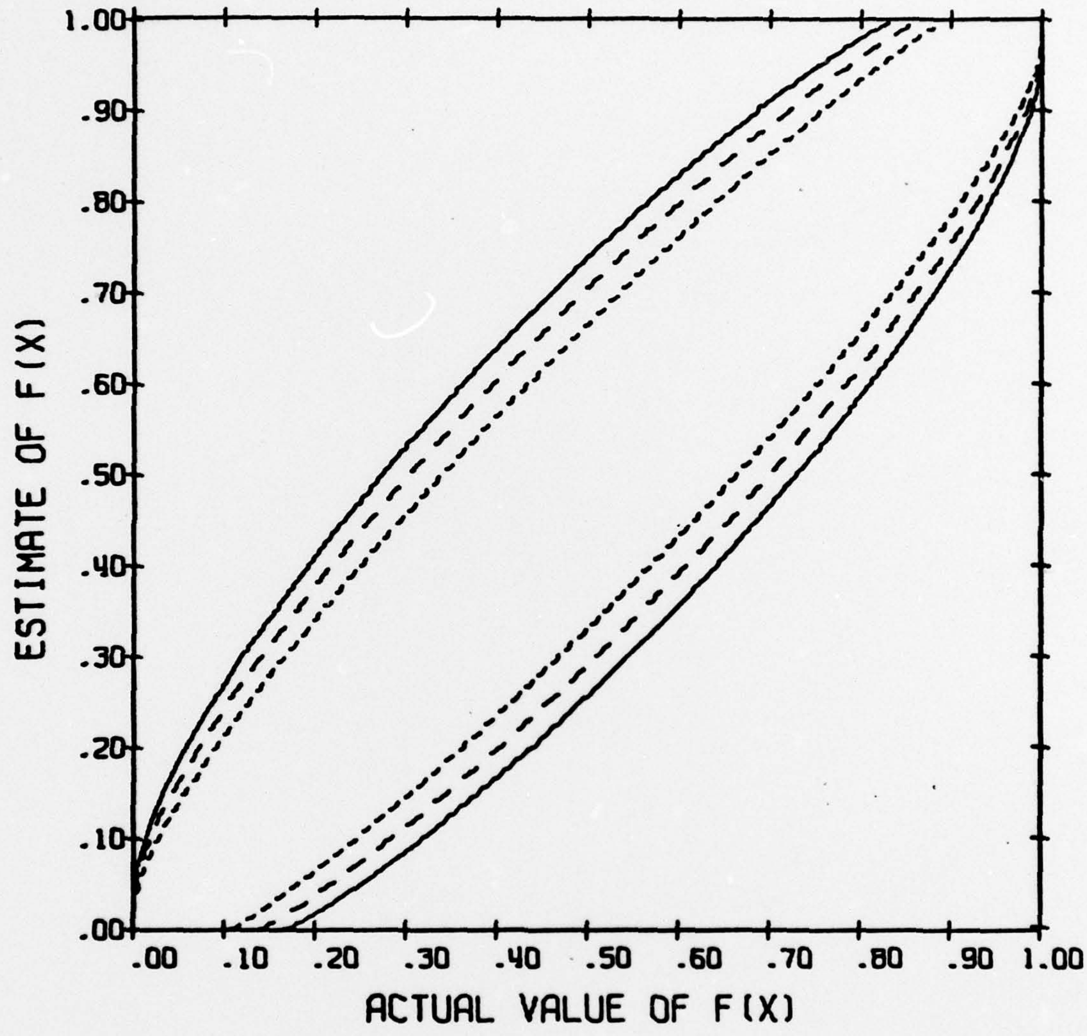


Fig. 6 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 20
——— .95 - - - .90 ----- .80

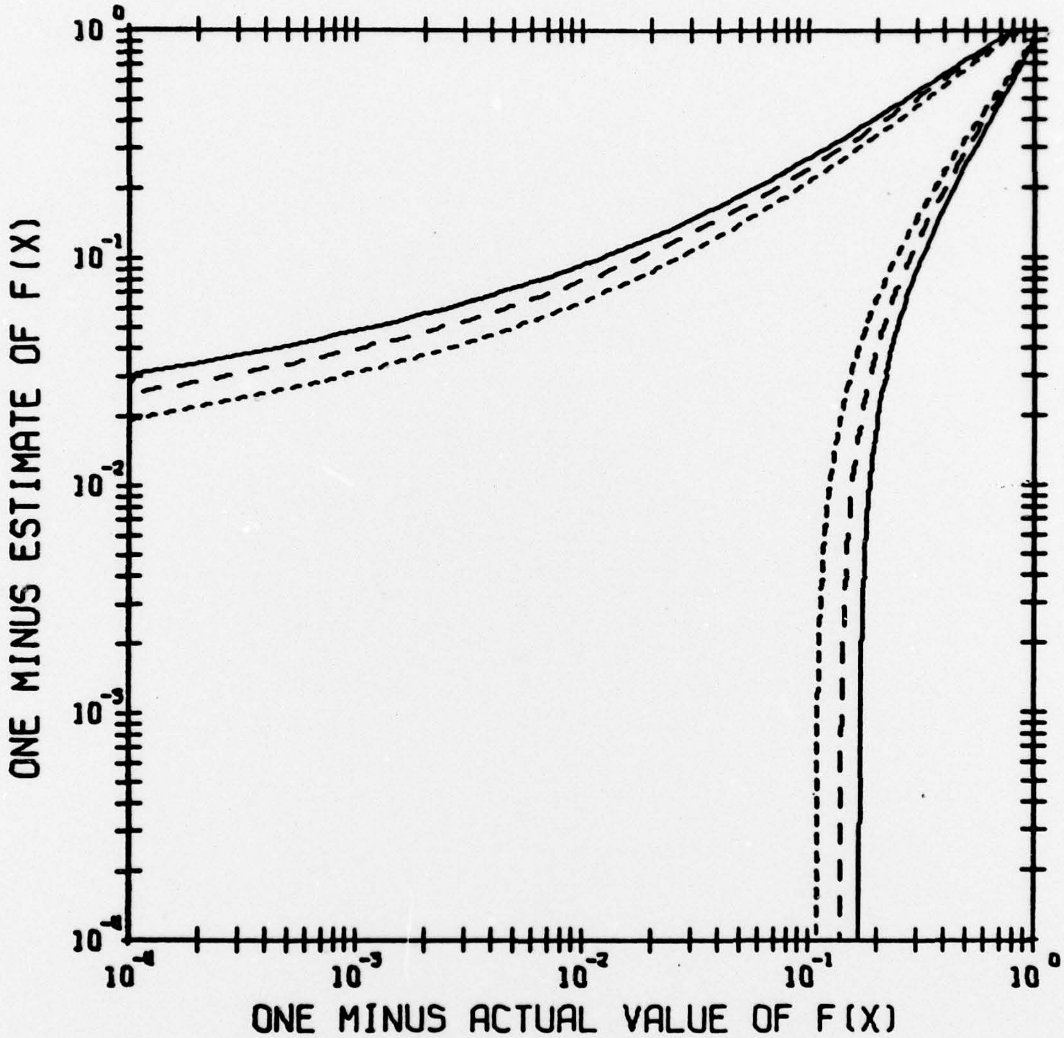


Fig. 7 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 20
——.95 - - -.90 -----.80

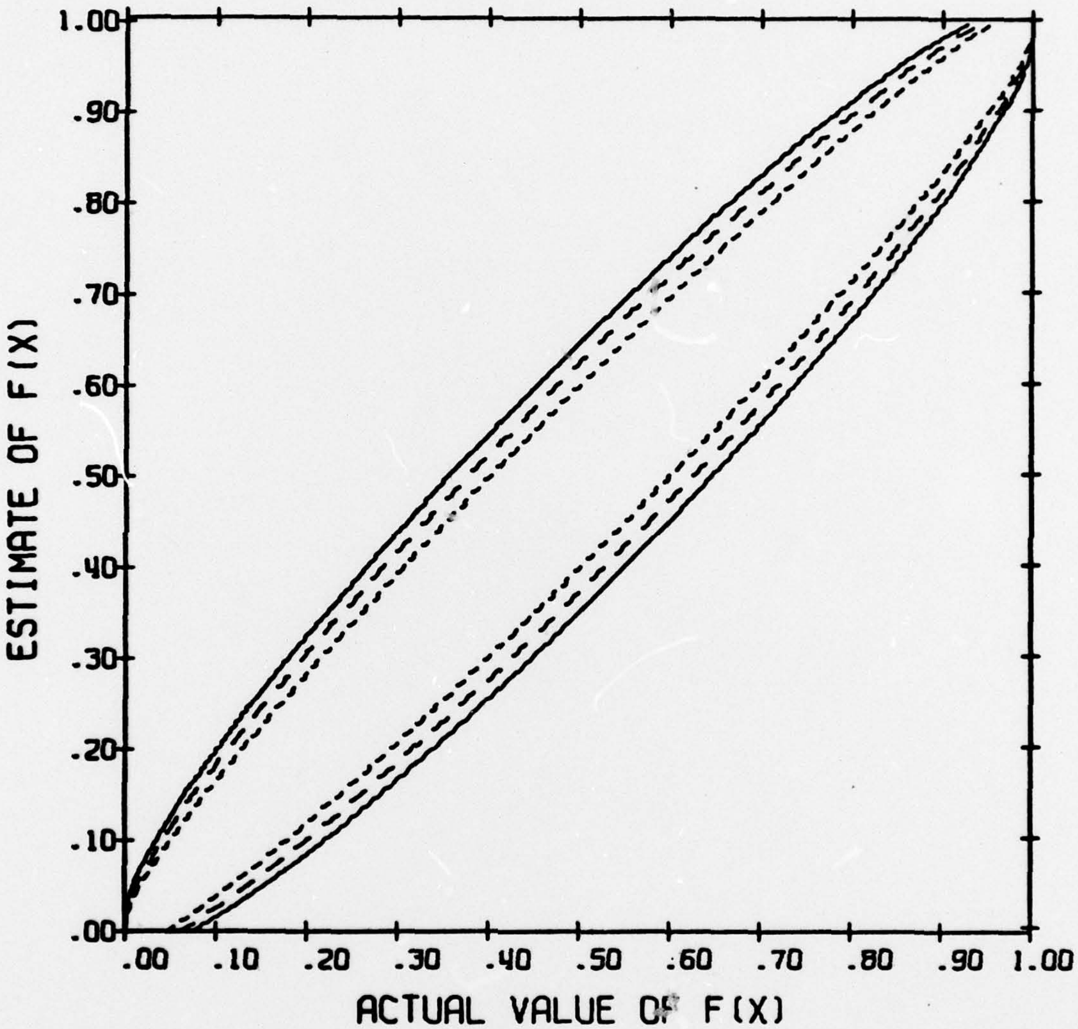


Fig. 8 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 50
——.95 - - -.90 -----.80

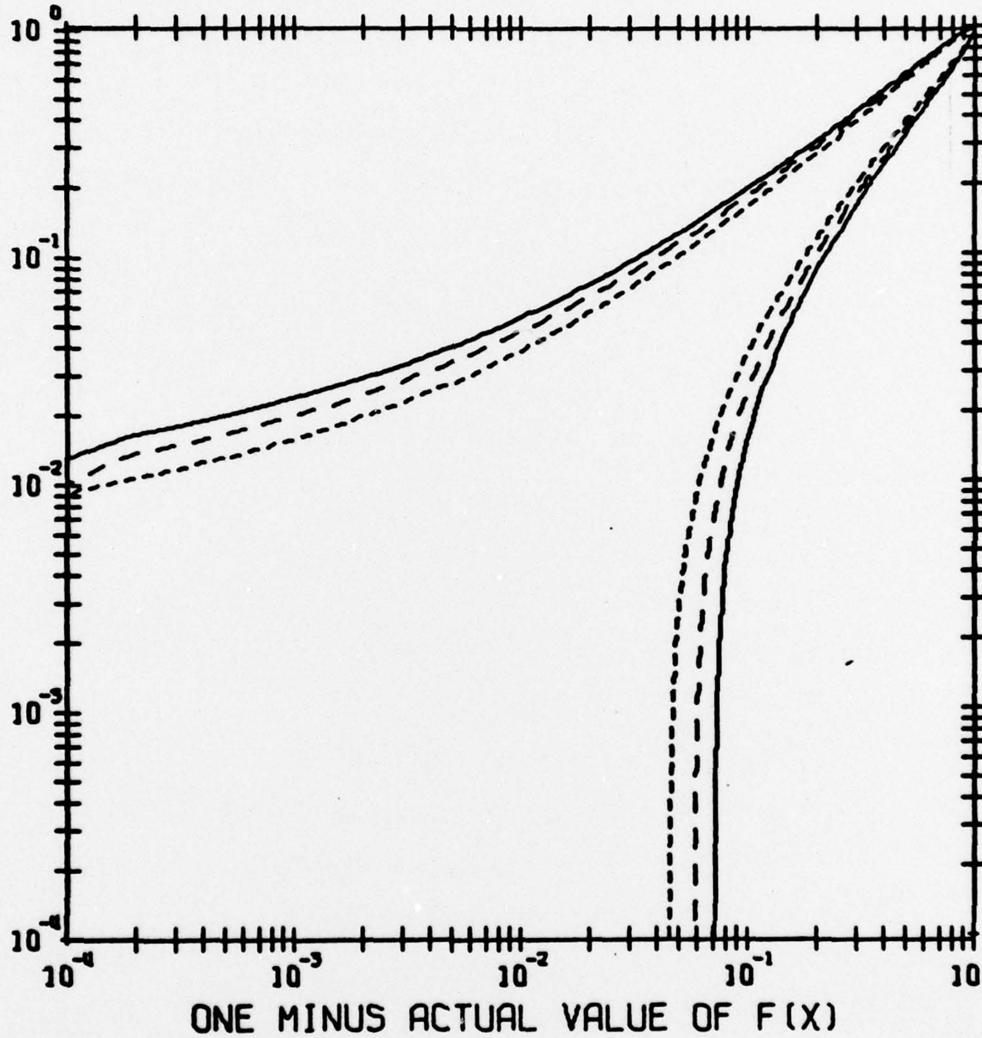


Fig. 9 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 50
——.95 - - -.90 -----.80

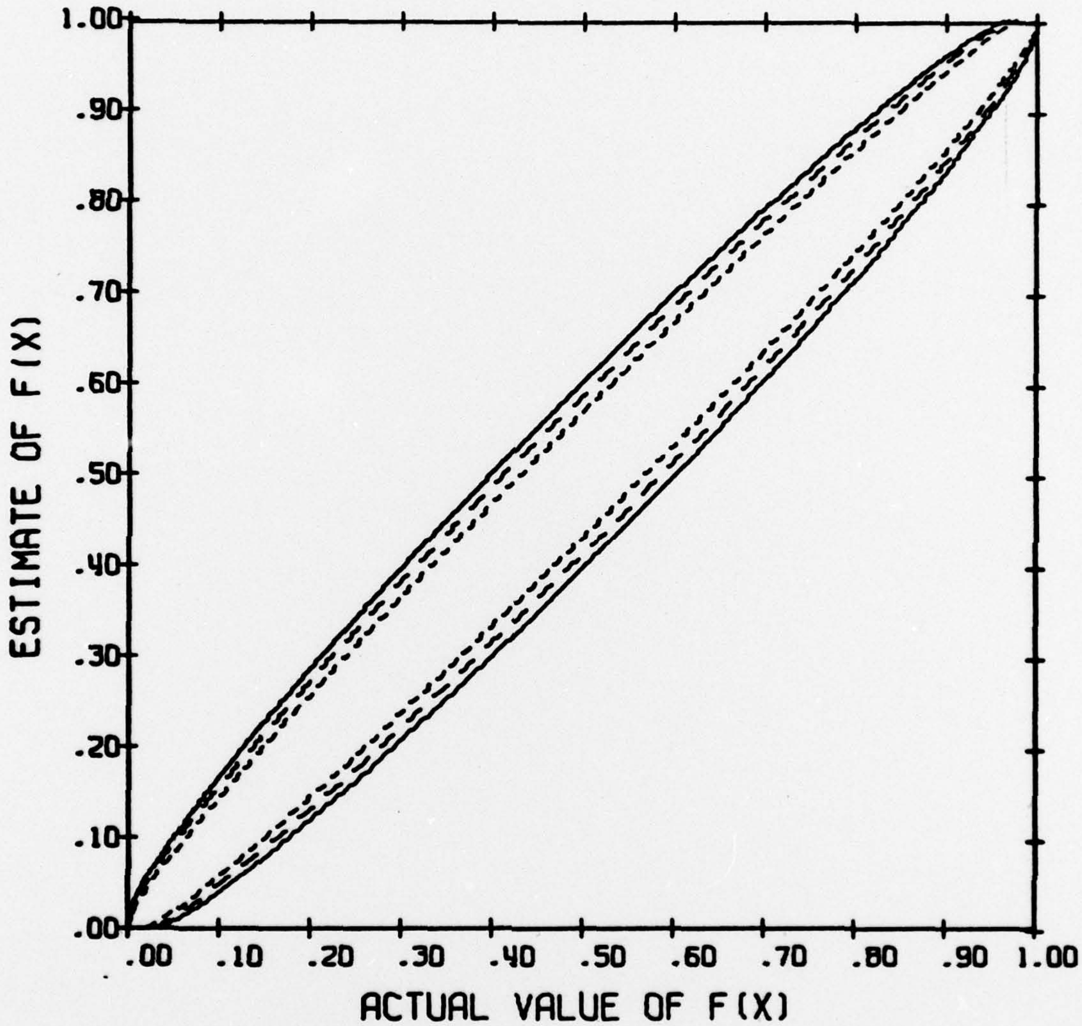


Fig. 10 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 100
——.95 - - -.90 -----.80

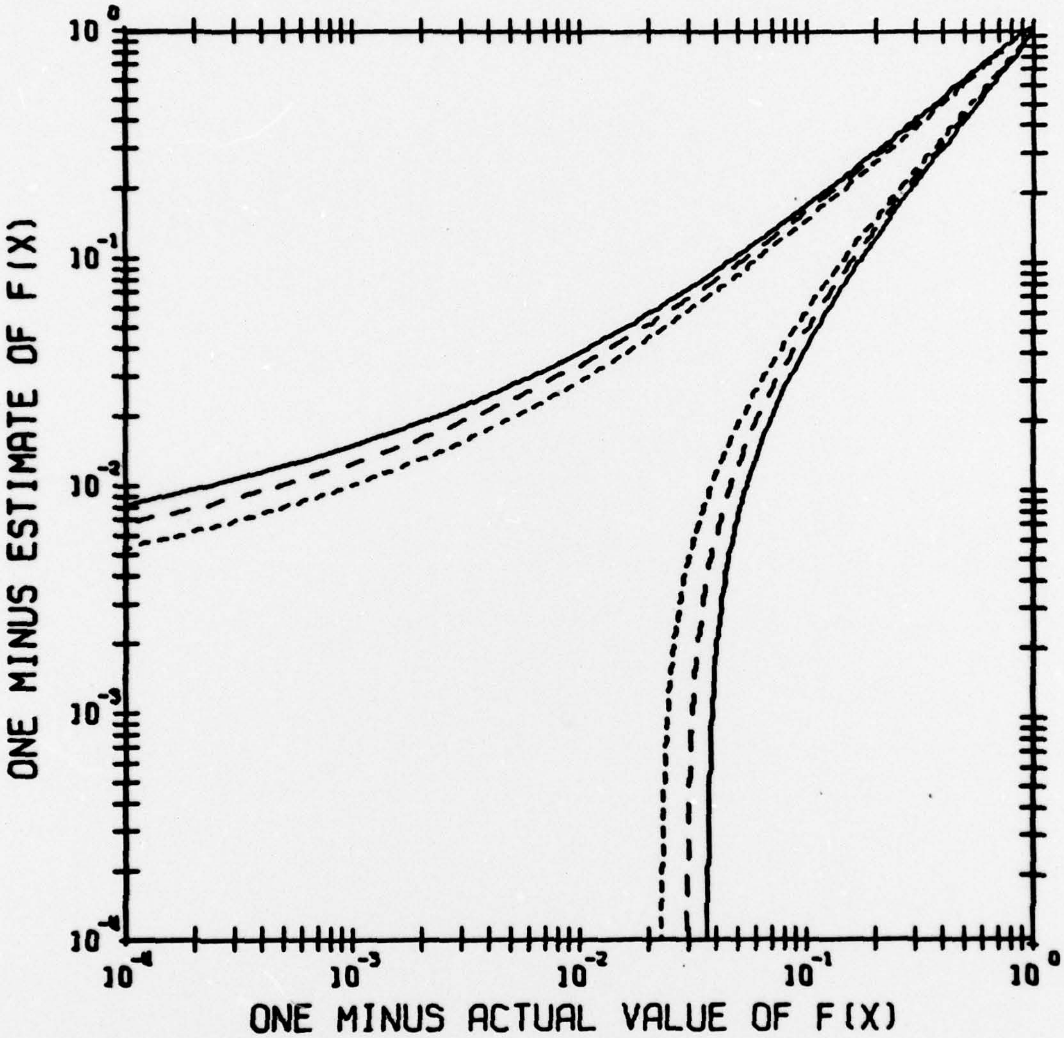


Fig. 11 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 100
——.95 - - - .90 -----.80

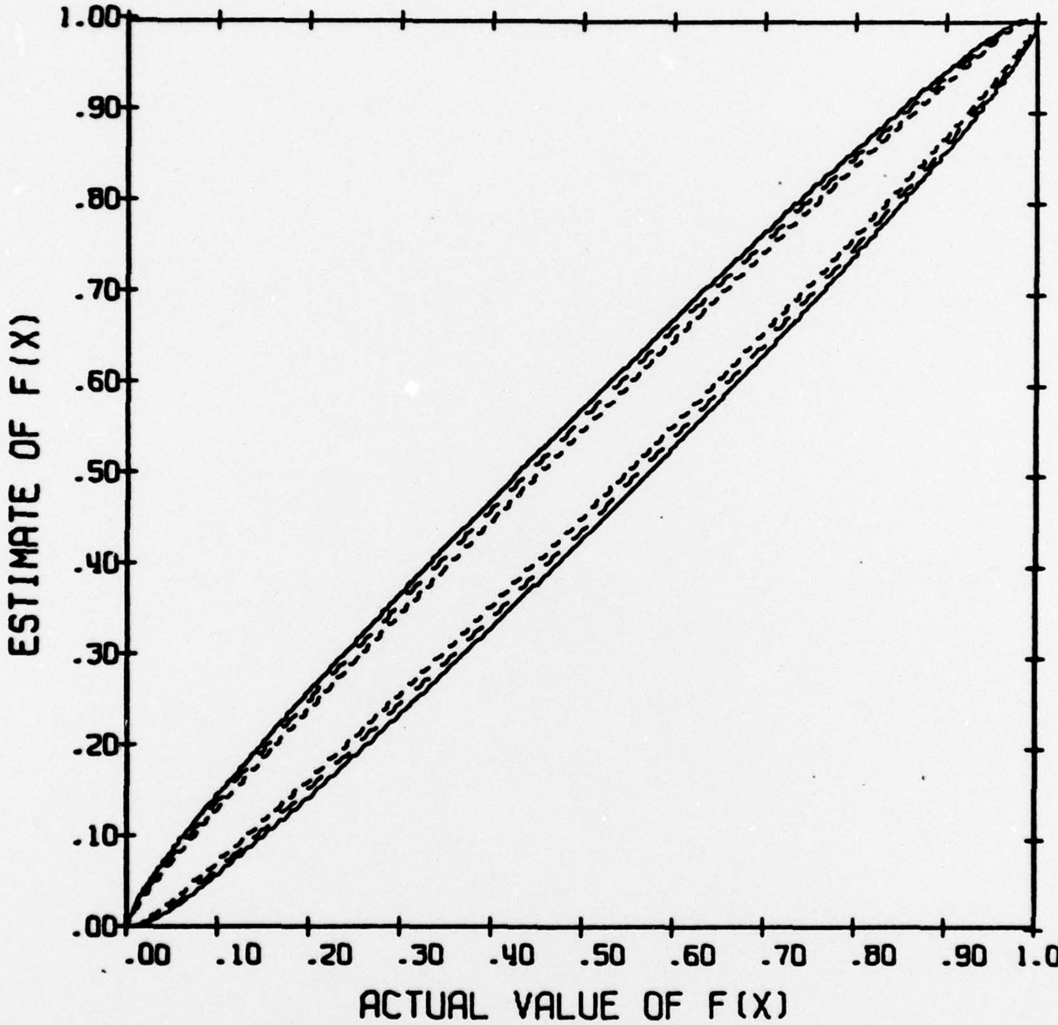


Fig. 12 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 200
——.95 - - -.90 -----.80

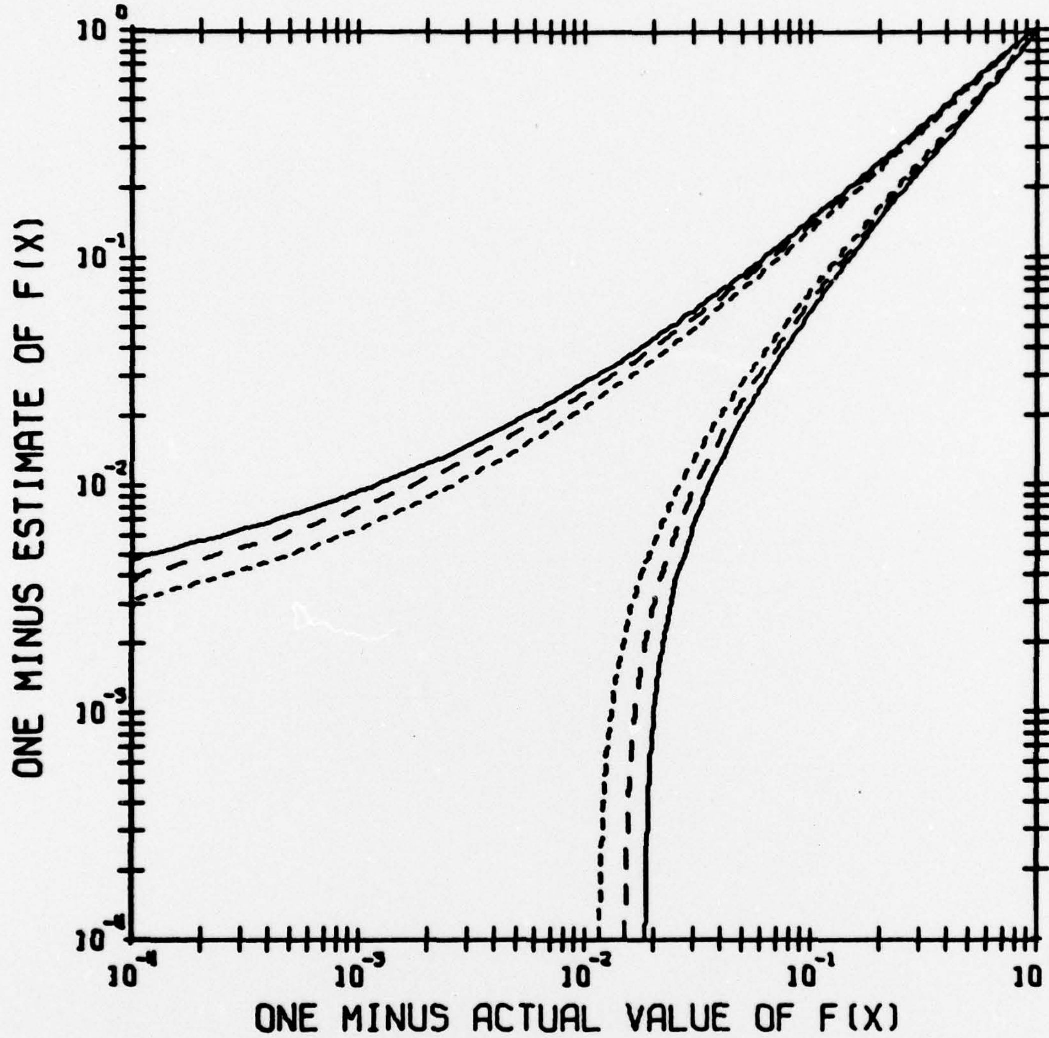


Fig. 13 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OFINDEPENDENT SAMPLES = 200
——.95 - - -.90 -----.80

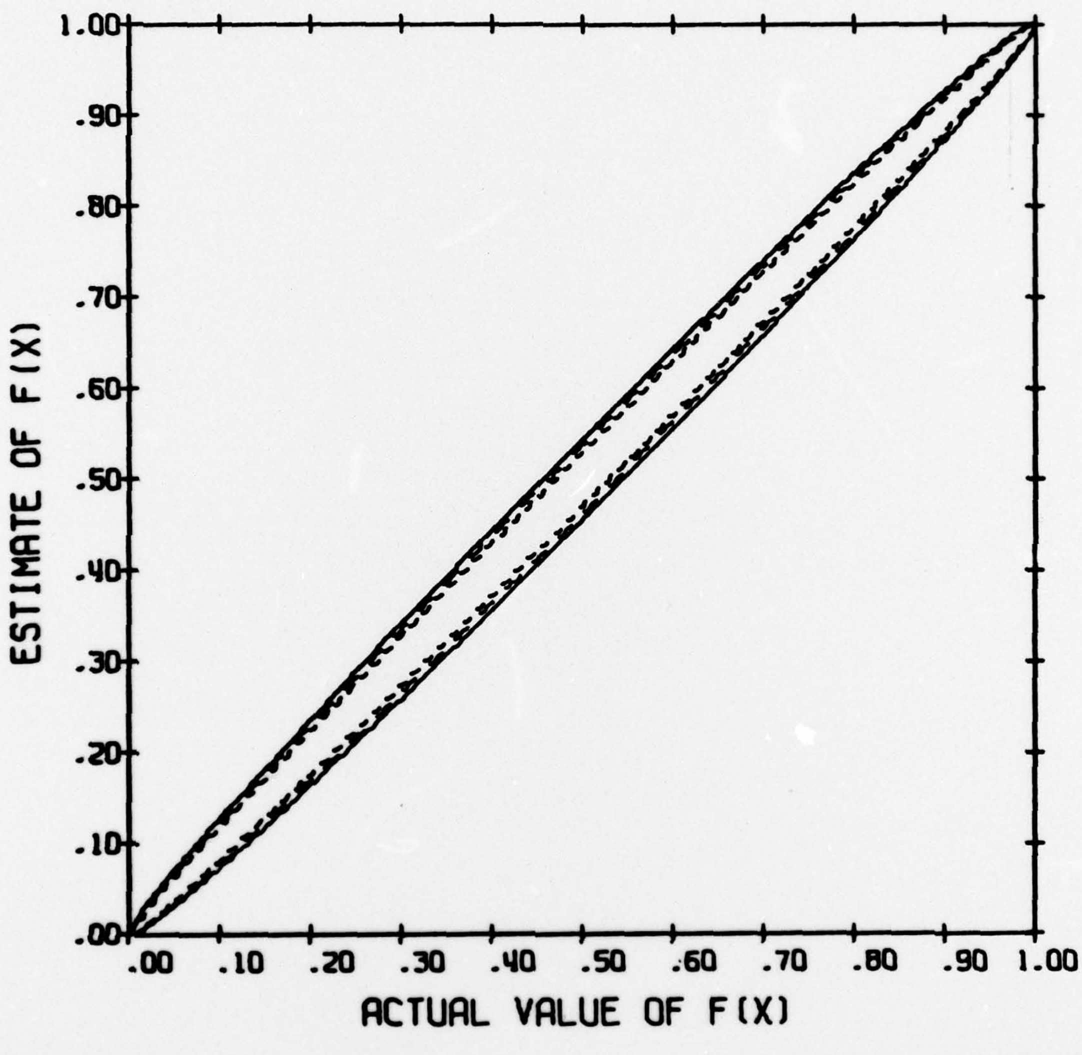


Fig. 14 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 500
——.95 - - -.90 -----.80

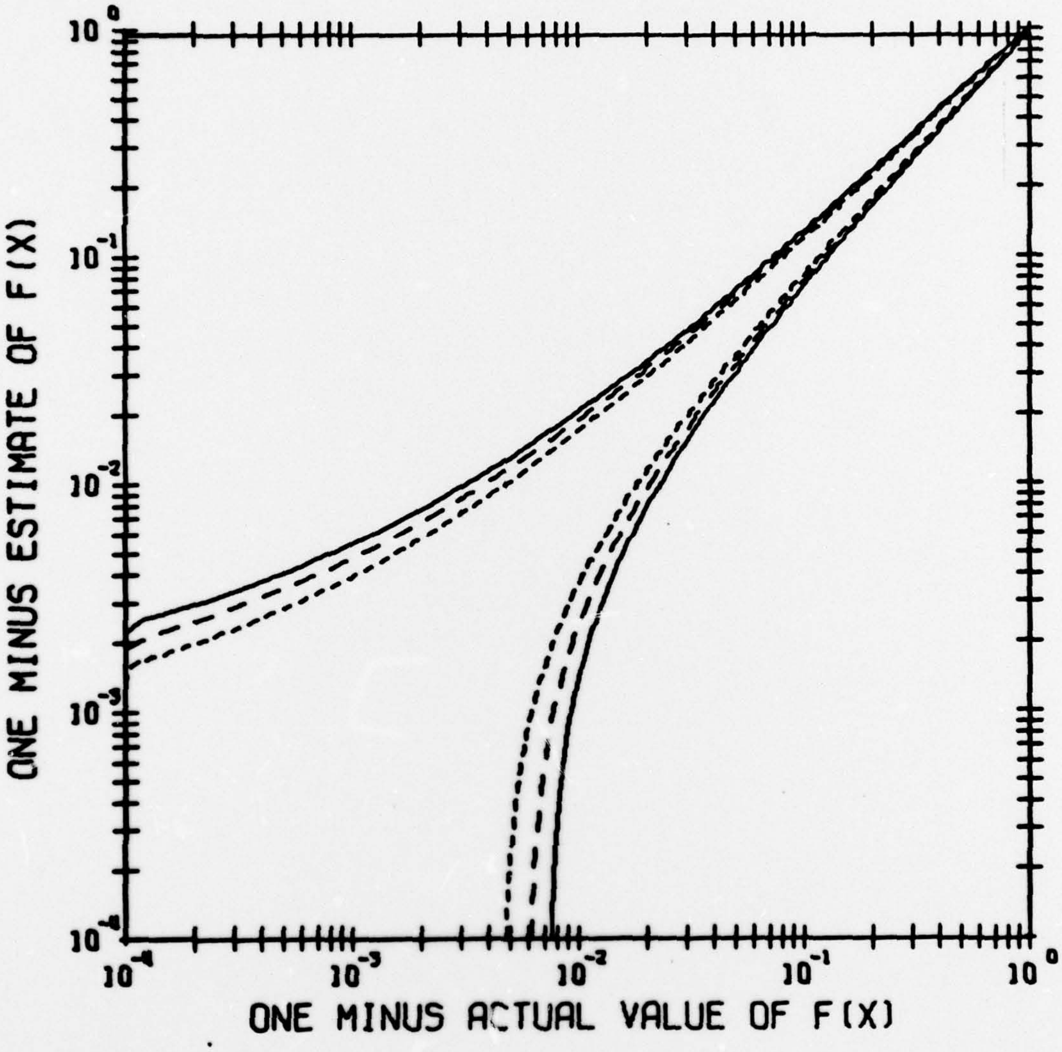


Fig. 15 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 500
——.95 - - -.90 -----.80

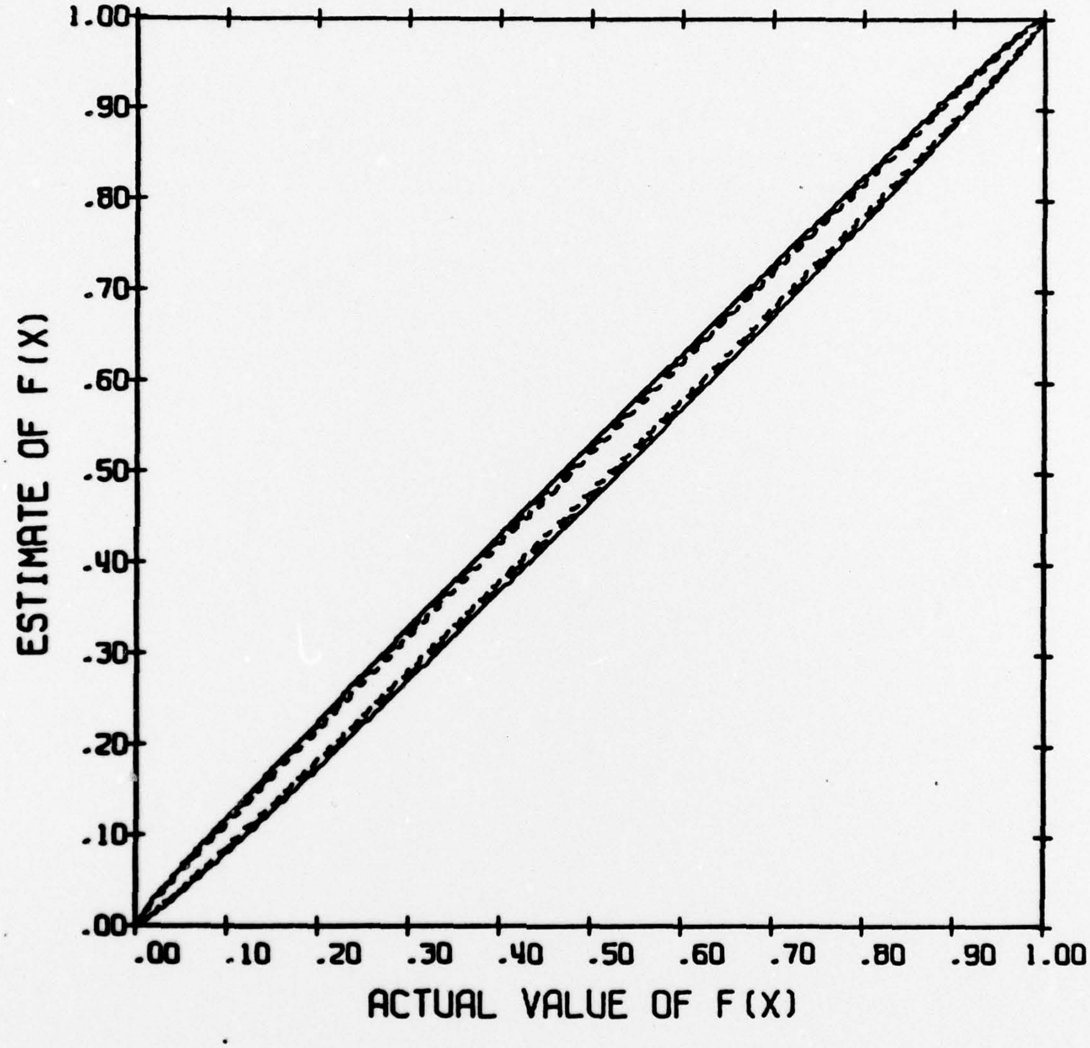


Fig. 16 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OFINDEPENDENT SAMPLES = 1000
——.95 - - -.90 -----.80

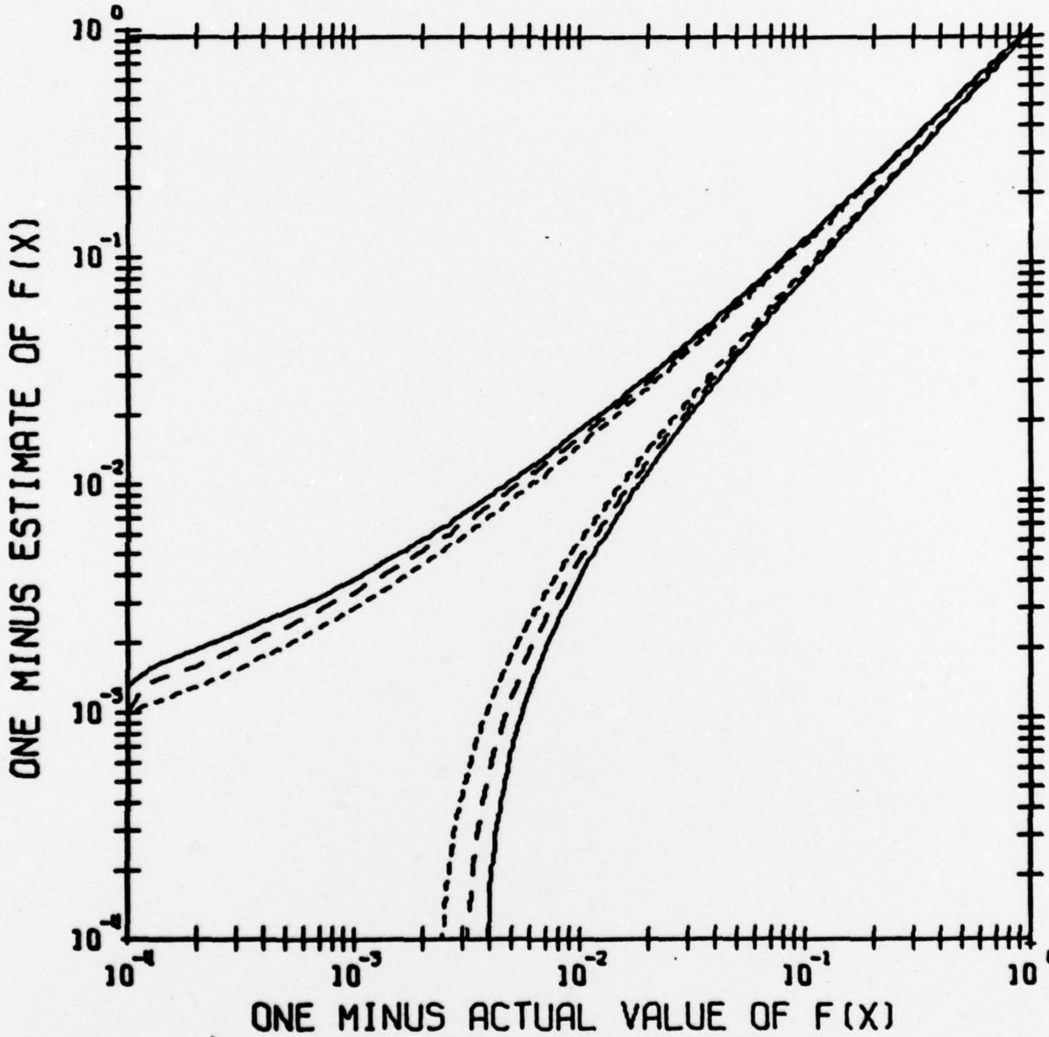


Fig. 17 CONFIDENCE INTERVALS FOR DISTRIBUTION FUNCTION
NUMBER OF INDEPENDENT SAMPLES = 1000
——.95 - - -.90 -----.80

TRACOR 6500 TRACOR LANE, AUSTIN, TEXAS 78721

5. REFERENCES

1. Bendat, J. S. and A. G. Piersol, "Measurement and Analysis of Random Data," John Wiley and Sons, Inc., New York, 1966.

2. Burrington, R. S., and D. G. May, "Handbook of Probability and Statistics with Tables," Handbook Publishers, Inc., Sandusky, Ohio, 1953.

3. Kendall, M. G. and Stuart, "The Advanced Theory of Statistics," Hafner, New York, 1967.